SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
MOLECULAR EVOLUTION
COMPARATIVE GENOMICS
EVOLUTIONARY BIOLOGY
FUNCTIONAL GENOMICS

Received
11 April 2012

Accepted
15 May 2012

Published
31 May 2012

Correspondence and requests for materials should be addressed to M.I. (mirimia@gmail.com) or J.G.F. (jordigarcia@ub.edu)

Comparative genomics of the *Hedgehog* loci in chordates and the origins of *Shh* regulatory novelties

Manuel Irimia[1,2], Jose L. Royo[3], Demian Burguera[1], Ignacio Maeso[1,4], José L. Gómez-Skarmeta[3] & Jordi Garcia-Fernandez[1]

[1]Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Barcelona, Spain, [2]The Donnelly Centre, University of Toronto, Toronto, Ontario, Canada, [3]Centro Andaluz de Biología del Desarrollo (CABD), Consejo Superior de Investigaciones Científicas/Universidad Pablo de Olavide, Sevilla, Spain, [4]Department of Zoology, University of Oxford, South Parks Road, Oxford, UK.

The origin and evolution of the complex regulatory landscapes of some vertebrate developmental genes, often spanning hundreds of Kbp and including neighboring genes, remain poorly understood. The *Sonic Hedgehog* (*Shh*) genomic regulatory block (GRB) is one of the best functionally characterized examples, with several discrete enhancers reported within its introns, vast upstream gene-free region and neighboring genes (*Lmbr1* and *Rnf32*). To investigate the origin and evolution of this GRB, we sequenced and characterized the *Hedgehog* (*Hh*) loci from three invertebrate chordate amphioxus species, which share several early expression domains with *Shh*. Using phylogenetic footprinting within and between chordate lineages, and reporter assays in zebrafish probing >30 Kbp of amphioxus *Hh*, we report large sequence and functional divergence between both groups. In addition, we show that the linkage of *Shh* to *Lmbr1* and *Rnf32*, necessary for the unique gnatostomate-specific *Shh* limb expression, is a vertebrate novelty occurred between the two whole-genome duplications.

How the extremely complex regulatory landscapes of certain developmental genes are originated and assembled in evolution is unclear. Although the presence of genomic regulatory blocks (GRBs) – in which key developmental factors are linked to bystander genes that contain regulatory information for the former – has been extensively described[1–3], the origin and evolution of such syntenic blocks, and their potential implications for organismal evolution is still poorly understood. One of the best characterized examples of a functional GRB involves *Sonic Hedgehog* (*Shh*)[4,5], a major morphogen in animal development[6,7]. *Shh* has been implicated in a wide variety of ontogenetic processes, such as the dorso-ventral (D–V)[8] and antero-posterior (A–P)[9] patterning of the developing central nervous system (CNS), the development of limbs[10], inner ear[11], digestive system[12], etc. Accordingly, *Shh* shows a remarkably complex expression pattern during development, comprising four major domains at early stages: CNS, notochord, epithelial sheet from the oral cavity to the hindgut, and limbs[13].

This complexity of developmental functions and expression patterns is paralleled at the genomic level. In mouse, *Shh* enhancers are scattered across a vast regulatory landscape spanning more than 850 Kbp, including its two introns, a gene desert of 729 Kbp in the upstream intergenic region and two upstream neighboring transcriptional units, the bystander genes *Lmbr1* and *Rnf32*. This region constitutes a GRB around *Shh* conserved in most vertebrate species[14], and comprises all *Shh* enhancers identified to date. A subset of these enhancers drives *Shh* expression to CNS domains conserved across jawed vertebrates (Figure 1). In the developing spinal cord, *Shh* is expressed all along the floor plate, and this expression is crucial for proper D–V patterning of the neural tube and the differentiation of specific cell populations[8]. In mouse, this expression is directed by two enhancers (*Shh* Floor Plate Enhancers, SPFE1 and SFPE2) that are located proximally upstream of the *Shh* coding region, and in the second intron, respectively[15]. Expression in the brain is more complex[16,17], and is controlled by at least four different enhancers. In particular, within the diencephalon, *Shh* expression shows a characteristic dorsal expansion from the basal plate: the core of the *Zona Limitans Intrathalamica* (ZLI). The ZLI is an important secondary organizer that regulates specific diencephalic fates through the action of *Shh*[9]. ZLI expression, together with those in the midbrain and caudal diencephalon are driven by the *Shh* Brain Enhancer 1 (SBE1), also located within the

second intron, as SFPE2[15]. The other three brain enhancers, SBE2-4, control more rostral expression domains and are located far upstream from the *Shh* coding sequence[18].

Expression to other developing tissues is also driven by specific enhancers, recognizable on the basis of sequence conservation across different vertebrate groups (i.e. as highly conserved non-coding regions, HCNRs). Limb expression is controlled by an enhancer located within the bystander gene *Lmbr1*, ~800 Kbp upstream of *Shh* in mouse (MFCS1, also called ZPA Regulatory Sequence (ZRS), Figure 1A)[5,19–21]. Similarly, expression to postpharyngeal linings is driven by an enhancer conserved from mammals to amphibians (MACS1), which is located within an intron of the *Rnf32* gene (Figure 1A)[4]; two other enhancers (MFCS4 and MRCS1) are also located near *Rnf32* and promote *Shh* transcriptional activation in more anterior linings. Finally, regarding notochord expression, an enhancer (SNE) has been identified upstream of mouse *Shh*, seemingly overlapping with SFPE1, and, although they have not been characterized, at least two notochord enhancers lie within the gene desert upstream of *Shh*, according to BAC screenings in mouse[18].

Shh regulation has also been extensively studied in zebrafish. Perhaps surprisingly, the scenario is quite different, although most of the proximal enhancers can be traced by sequence similarity. Three HCNRs were identified within *shha* (two in intron 1, ar-A and ar-B, and one in intron 2, ar-C, Figure 1B), plus a forth HCNR upstream, near the transcription start site (ar-D) (Figure 1B). Enhancers ar-D and ar-C correspond to SFPE1 and SFPE2, respectively. Their function, however, differs from the mouse counterparts, which drive expression throughout the floor plate: ar-D drives expression only to the anterior floor plate, and ar-C promotes expression in forebrain and notochord, and only weakly in the floor plate[22]. On the other hand, ar-B drives expression throughout the spinal cord floor plate[22], and it has been lost in mammals[23], and ar-A drives expression to notochord, and some brain structures, similar to ar-C[22]. Phylogenetic footprinting using coelacanth – a slow-evolving, sister species of the tetrapods – show that these four HCNRs are ancestral; nonetheless, the enhancer function of the coelacanth sequences is more similar to the tetrapod counterparts[23]. The enhancer(s) responsible for other expression domains have not been characterized yet in zebrafish, although HCNRs orthologous to some of the mouse elements are present in teleost species[14,18,20,24].

Despite the fact that *Shh* seems to have taken most ancestral *Hedgehog* functions[6], tetrapods have two other paralogs, *Indian hedgehog* (*Ihh*) and *Dessert hedgehog* (*Dhh*), originated in the two rounds of whole genome duplication (WGD) occurred at the base of vertebrates[25]. The coding sequences of these paralogs are more divergent, and their developmental expression domains and functions are much more restricted than those of *Shh*, especially in the case of *Dhh*[6]. Accordingly, the regulation of both *Dhh* and *Ihh* have received little attention, and only one enhancer, responsible for the *Ihh*-specific expression during endochondral bone formation, has been identified so far[26]. This element is located within the longest intron of the upstream neighboring gene, *Nhej1*, suggesting that this gene is part of the *Ihh* GRB. In invertebrates, *Hedgehog* genes also show complex expression patterns and play crucial roles during development in all studied species[6,27–32]. In the basal chordate amphioxus, the best living proxy to the vertebrate-invertebrate ancestor bodyplan, *Hh* is expressed in four major developing regions at early developmental stages: CNS, notochord, tail bud and pharyngeal endoderm (including forming gill slits)[32,33], some of which readily correspond to vertebrate *Shh* expression domains. In the developing CNS, amphioxus *Hh* is also restricted to the ventral side of the forming neural tube up to a rostral limit; however, in stark contrast to all vertebrates, no expression is found in the most anterior part of the amphioxus CNS, including no dorsal ZLI-like expansion[32–35]. This suggests important changes in the regulation of *Shh/Hh* during chordate

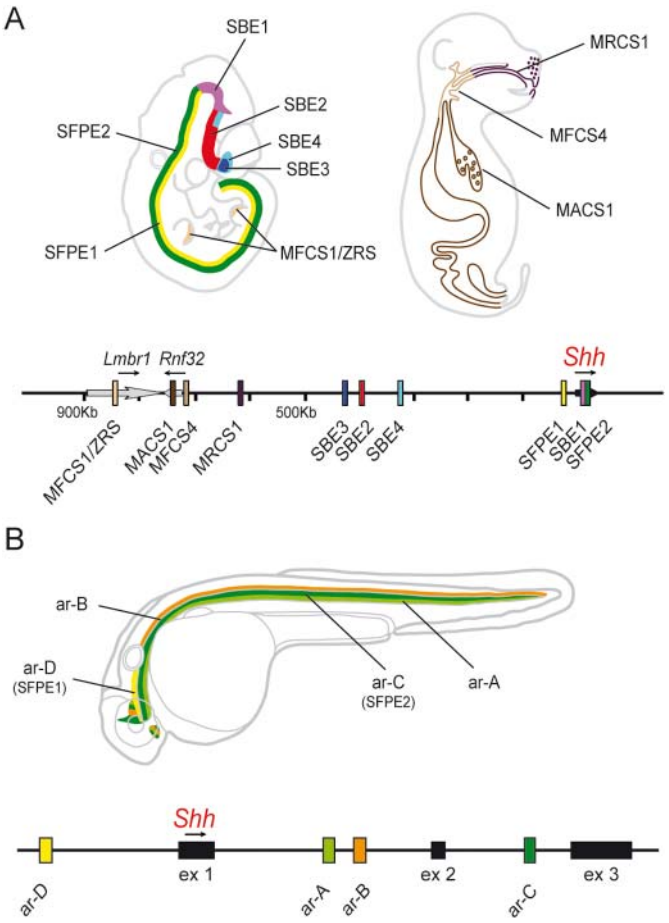

Figure 1 | Genomic location of tissue-specific *Shh* enhancers in mouse and zebrafish. (A) Distribution of tissue-specific enhancers across the large upstream region and introns of *Shh* in mouse chromosome 5. Each enhancer is represented as a color block and its associated expression is shown in the same color in the schematic embryos above. SFPE1 (green) and SFPE2 (yellow) drive expression throughout the floor plate of the spinal cord; SBE1 (lile), to the midbrain and caudal diencephalon, including the ZLI; SBE2-4 (red and dark and light blue), to more anterior domains in the developing brain; MRCS1 (purple), MFCS4 (light brown) and MACS1 (dark brown) to epithelial linings; and MFCS1/ZRS (light orange) to limb buds. Two enhancers lay within the intronic sequence of bystander genes *Lmbr1* (*MFCS1/ZRS*) and Rnf32 (*MACS1*), and two within the second intron of *Shh* (SBE1 and SFPE2). (B) Distribution of known enhancers in zebrafish *shha* gene. ar-A (light green) drives expression to the notochord and some brain structures; ar-B (dark orange), throughout the spinal cord floor plate; ar-C (dark green), to forebrain and notochord, and weakly in the floor plate; and ar-D (yellow), to the anterior floor plate. Adapted from different sources[4,18,22].

diversification; however, the evolution of *Hh* regulatory landscape is still poorly understood.

Here, we have analyzed the amphioxus *Hh* genomic locus to get insights into the origin and evolution of the vertebrate *Shh* GRB. We have sequenced ~55 Kbp of *Hh* loci in the European amphioxus, *Branchiostoma lanceolatum,* and performed phylogenetic footprinting analyses with two sister species (the Floridian and Chinese amphioxus), and several vertebrates. We found widespread conservation of non-coding sequences within the amphioxus *Hh* locus between the three cephalochordates, but we could not identify reliable orthologous sequences to any of the vertebrate HCNRs. In order to test cryptic regulatory conservation, we also generated transgenic zebrafish lines carrying amphioxus sequences spanning the whole

Hh locus. One region drove reporter expression consistent with the endogenous *Hedgehog* genes in zebrafish and amphioxus (in developing pharyngeal endoderm and gill slits). This sequence and *cis*-regulatory function has no evolutionary correspondence to any described vertebrate enhancer, further supporting a general lack of regulatory conservation between vertebrates and amphioxus. Finally, we investigated the microsynteny associated with the *Hedgehog* locus in vertebrates and invertebrates. We found strong evidence for a vertebrate-specific genomic rearrangement affecting *Shh/Dhh* between the two rounds of WGD that configured a novel microsyntenic environment that included the enhancer-containing bystander genes *Lmbr1* and *Rnf32* as parts of a new GRB.

Results

Cloning and sequencing of *B. lanceolatum Hh* loci and comparison to other amphioxus species. We sequenced ~55 Kbp of genomic sequence of the *Hh* locus from the European amphioxus, *B. lanceolatum*, following two main strategies (see Methods). A genomic fragment of 41,161 bp, containing the three coding exons of *Hh*, was sequenced from different phages identified through screening of a genomic library. In addition, 13,640 bp further upstream from the *Hh* loci were cloned using primers designed on lowly polymorphic *B. floridae* regions, based on haplotype comparisons. In total, the sequenced region comprised 29,069 bp upstream of the start codon, the three coding exons, the first and second introns (11,458 and 10,312 bp, respectively), and 2,714 bp downstream of the termination codon, including the full 3′ untranslated region (Figure 2A).

We next compared this sequence with the orthologous genomic regions from the Floridian and Chinese amphioxus species, *Branchiostoma floridae* and *Branchiostoma belcheri*, using LAGAN alignments visualized as VISTA plots (Figure 2B). Despite the three amphioxus species diverged at least 100 million years (my) ago[36–38], we found widespread conservation of non-coding sequences, even using highly stringent conditions (calculation window=300 bp, minimum width=300 bp, sequence identity=80%). The conservation of non-coding sequences was particularly striking within the two *Hh* introns, with regions having sequence similarity of 90% over >1,500 bp among the three amphioxus species (Figure 2B).

Comparison of *Hedgehog* loci from amphioxus and vertebrates. We next compared the amphioxus *Hh* locus with the vertebrate paralogs, *Shh*, *Dhh* and *Ihh*. One of the main differences between these loci is the massive upstream gene-free region in most *Shh* genes, compared to amphioxus *Hh* and the other two vertebrate paralogs. For instance, in mouse, the region upstream of *Shh* up to *Rnf32* comprises 729 Kbp, whereas *Ihh* and *Dhh* have upstream intergenic regions of 16 and 5.4 Kbp, respectively, and *B. belcheri Hh* has 27 Kbp. These differences suggest higher complexity in the regulatory landscapes for the *Shh* genes. On the other hand, the two conserved introns are more than twice the length in amphioxus *Hh* than in *Shh* genes, despite the fact that several enhancers have been described within these vertebrate introns[15,22]. Thus, it could be possible that much of the *cis*-regulatory information in amphioxus is also contained within these introns.

We attempted to identify deeply conserved HCNRs across chordates using VISTA plots of different alignment software for amphioxus *Hh* and several vertebrate *Shh* loci (Figure 3, see Methods). Several vertebrate- or tetrapod-specific HCNRs were identified in the *Shh* upstream region (Figure 3A), including all previously characterized enhancer elements[4,23]. However, none of these elements seems to be significantly conserved in amphioxus, even using highly relaxed conservation parameters (see Methods). For example, using LAGAN alignments, VISTA analysis detected a possible trace of conservation only for SBE4. (Figure 3A, light blue); nonetheless, this short sequence had low complexity and was not conserved between the three amphioxus species (see also Figure S1), despite their

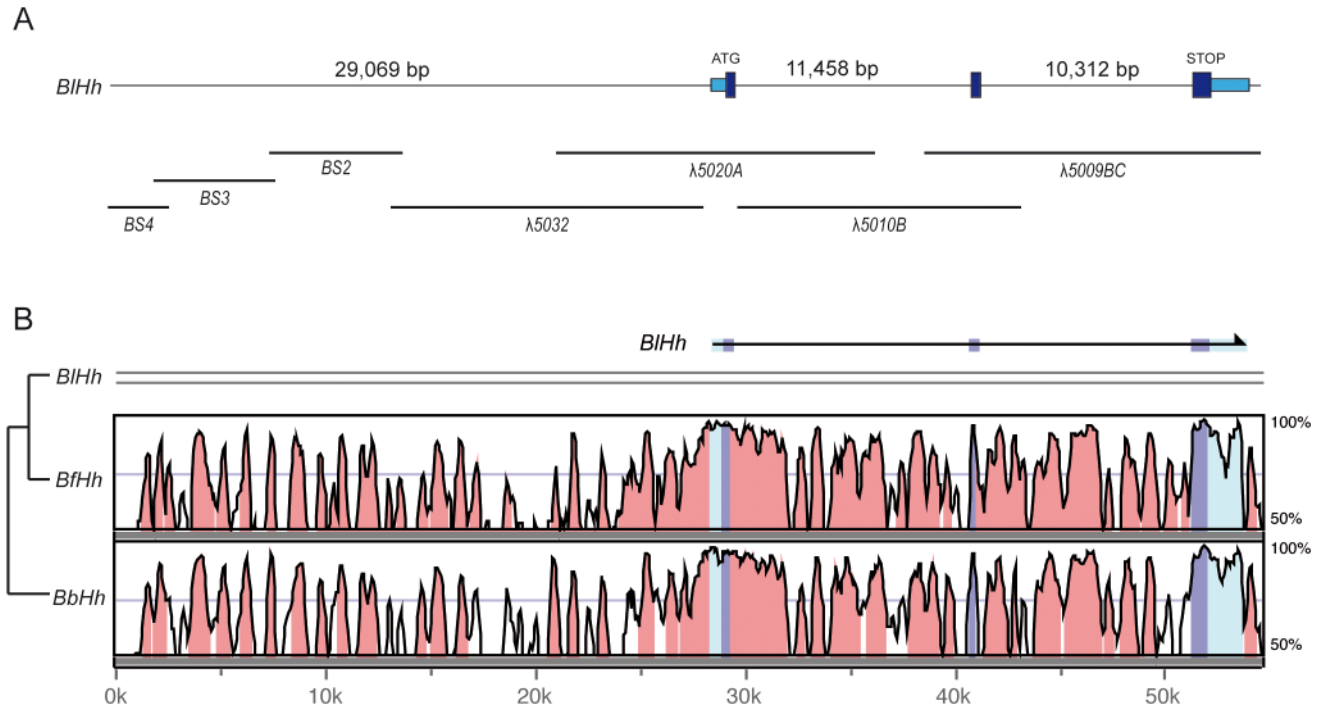

Figure 2 | Comparison of *Hh* locus among three amphioxus species. (A) Schematic representation of the *B. lanceolatum Hh* genomic region and genomic fragments cloned and sequenced in this study. Coding exons are shown in dark blue and UTRs in light blue. Phage genomic fragments are named after λ, and A–C indicates the exon(s) contained in the sequence. (B) VISTA plot comparing *B. lanceolatum* (reference, on top) with *B. floridae* (*BfHh*) and *B. belcheri* (*BbHh*) *Hh* loci using highly stringent conditions (LAGAN alignment, window size=300 bp, minimum width=300 bp, identity=80%). Dark/light blue indicates coding/UTR exonic sequence and pink shows non-coding regions conserved above threshold.

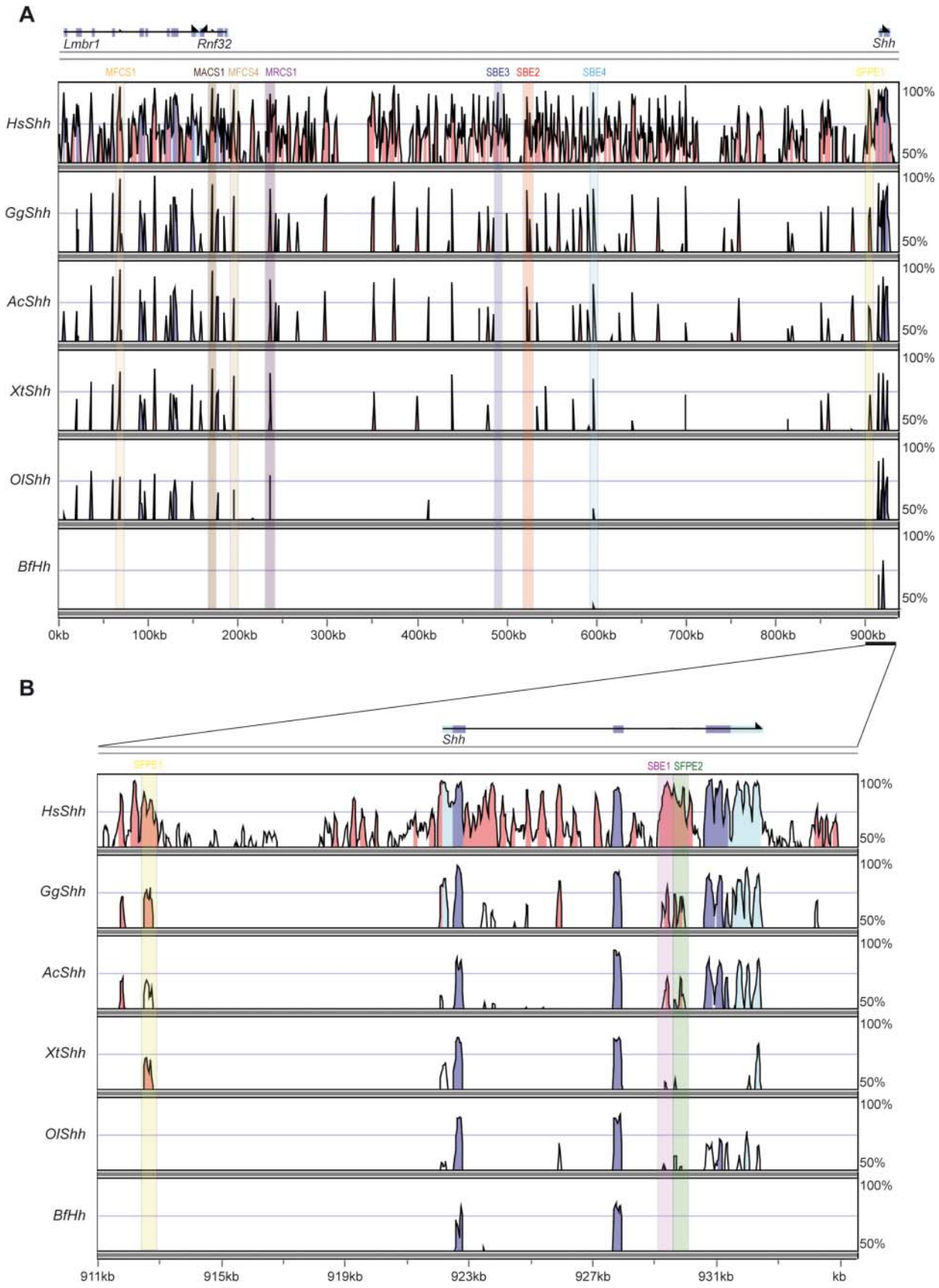

Figure 3 | Comparison of amphioxus *Hh* and vertebrate *Shh* loci. (A) VISTA plot for the alignment of the *Shh* genomic region from six vertebrate species and amphioxus *Hh* using mouse *Shh* as reference with default conditions, except for amphioxus (window size of 50 bp, minimum width of 50 bp and sequence identity threshold of 60%). The aligned sequences comprise the genomic region between *Lmbr1* and *Shh*, both included. (B) Detailed alignment within the *Shh* loci, as indicated in (A). Each reported tissue-specific enhancer is highlighted using the same color code as in Figure 1. Species abbreviations: *HsShh*, human, *GgShh*, chicken, *AcShh*, green anole lizard, *XtShh*, *Xenopus tropicalis*, *OlShh*, medaka, and *BfHh*, Floridian amphioxus.

widespread general sequence identity, suggesting it is likely a false positive.

Similarly, no conservation of non-coding regions was observed within the two *Shh* introns and for the sequence corresponding to the proximal floor plate enhancer SFPE1 (Figure 3B). Given that the two SFPE enhancers drive *Shh* expression to specific CNS subdomains that are likely homologous to those of amphioxus developing neural tube[32], we performed specific local alignment of these sequences to the corresponding amphioxus regions. No trace of sequence conservation was found for any of the enhancers, including full sequence motif arrangements for previously described functional transcription factor binding sites[39]. Consistently, only a partial arrangement for SFPE2, with a putative FoxA2 binding site, has been previously reported[40]. Finally, comparison of amphioxus *Hh* to the other two vertebrate paralogs, *Dhh* and *Ihh*, gave similar negative results (Figure S2). Within jawed vertebrates, however, a few HCNRs were detected. In the case of *Ihh*, they were restricted to tetrapods (most of them to amniotes) and mostly located within the introns of its upstream neighboring gene, *Nhej1*, consistent with previous results[26]. *Dhh* presented a more extreme situation and no conservation could be identified outside mammals, in line with this paralog having a much simpler transcriptional regulation than *Shh*. Using amphioxus, medaka or Xenopus as reference genomes for the above analyses yielded similar results (data not shown).

Enhancer activity of amphioxus *Hh* sequences in transgenic assays in zebrafish. The lack of non-coding sequence conservation over long phylogenetic distances is not particularly surprising, since it is known to be rare[3,41,42]. However, despite lack of sequence similarity, positive enhancer activity from amphioxus sequences has been successfully detected in zebrafish transgenic assays[3], presumably reflecting conservation of ancestral chordate regulatory states. To investigate if this was the case for *Hedgehog*, we assayed >30 Kbp from the amphioxus *Hh* locus for enhancer activity using zebrafish transgenesis. Since the widespread conservation of non-coding regions among the amphioxus species precluded the identification of discrete candidate HCNRs, we generated zebrafish lines carrying overlapping fragments spanning both introns and ~11 Kbp upstream from the transcription start site (Figure 4A). Only two fragments (D and F) drove consistent mosaic reporter expression at 24 hpf or 48 hpf embryos, but only F, within the second intron, was consistent with the endogenous *shh* expression (D drove expression to the hatching gland, Figure S3). To better determine the enhancer activity within the region F, we generated stable transgenic lines for this fragment. Three out of four different stable lines showed GFP expression in the developing pharynx and gill slits, confirming the results from the F0 assays. In situ hybridization of GFP transcripts confirmed reporter expression to developing pharyngeal endoderm and branchial arches, but not in notochord or CNS (Figure 4D, and transversal section in Figure 4E). This expression is part of the endogenous expression pattern of zebrafish *shh* genes (arrow head in Figure 4C[43],), and presumably homologous to the expression of amphioxus *Hh* in developing pharyngeal endoderm and gills slits[33]. In addition, we also generated stable transgenic zebrafish lines for fragments B and G, spanning the equivalent regions to those where the floor plate enhancers lay in vertebrates (Figure 1). None of these regions activated GFP expression in the transgenic embryos at these stages, and only founders with control RFP expression were identified for each construct (data not shown).

Synteny analysis of vertebrate and invertebrate *Hedgehog* loci identifies a genomic rearrangement that has remodeled Shh regulatory landscape. To reconstruct the evolutionary history of *Shh* GRB, we studied the local synteny surrounding members of the *Hedgehog* gene family across metazoans (Figure 5). Within jawed vertebrates, we found a clear correspondence for general genetic neighborhood for the three vertebrate *Hedgehog* genes, with the region upstream of each *Hedgehog* gene containing at least one gene from three gene families (*Des/Prph*, *DnaJB* and *Tub*) (Figure 5A). This pattern suggests that this cluster of genes is an ancestral local linkage group, established before the WGDs that gave rise to the three *Hedgehog* paralogs in vertebrates. Interestingly, the genes immediately adjacent to the *Hh* paralogs showed a more complex pattern. As mentioned above, *Shh* is neighbored by the upstream bystander genes *Lmbr1* and *Rnf32*, which contain important regulatory elements for *Shh*; however *Dhh* contains only one of these genes (the paralog *Lmbr1l*), and

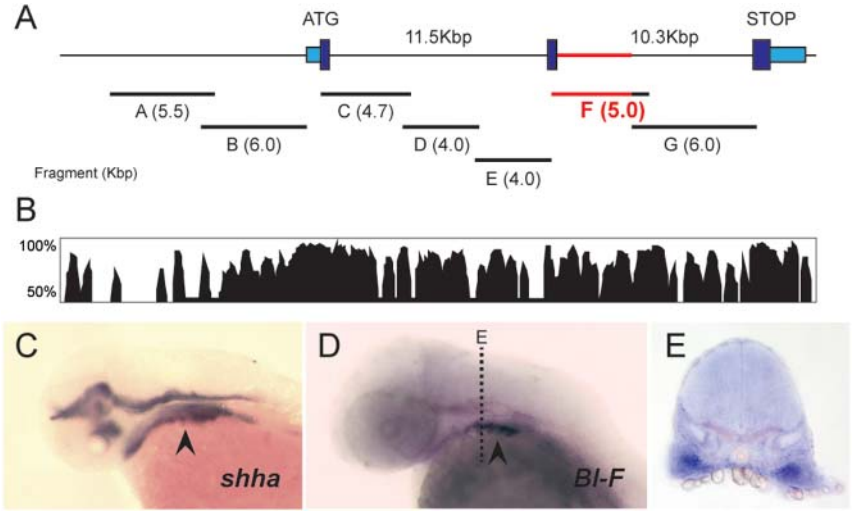

Figure 4 | Transgenic reporter analysis of amphioxus *Hh* sequences in zebrafish. (A) Schematic representation of the location and length of the seven fragments (A–G) spanning >30 Kbp of the amphioxus *Hh* locus tested by transgenesis in zebrafish. In red, 'F', the only fragment that drove GFP reporter expression consistent with the endogenous genes in zebrafish and amphioxus (D–E). (B) Conservation of the *B. lanceolatum* sequence compared to *B. floridae*, as in Figure 1. (C) *In situ* hybridization of *shha* in a zebrafish 30 hpf embryo. Arrowhead shows expression in pharyngeal endoderm and forming gill slits. (D) *In situ* hybridization of GFP in a stable transgenic embryo carrying fragment F. Expression is only observed in pharyngeal endoderm and gill slits (note that the seemingly dorsal expression domain correspond to the expression of the opposite side, as shown in (E)). (E) Section through dashed line in (D) showing expression of GFP is restricted to pharyngeal endoderm and gill slits, and not present in notochord or CNS.

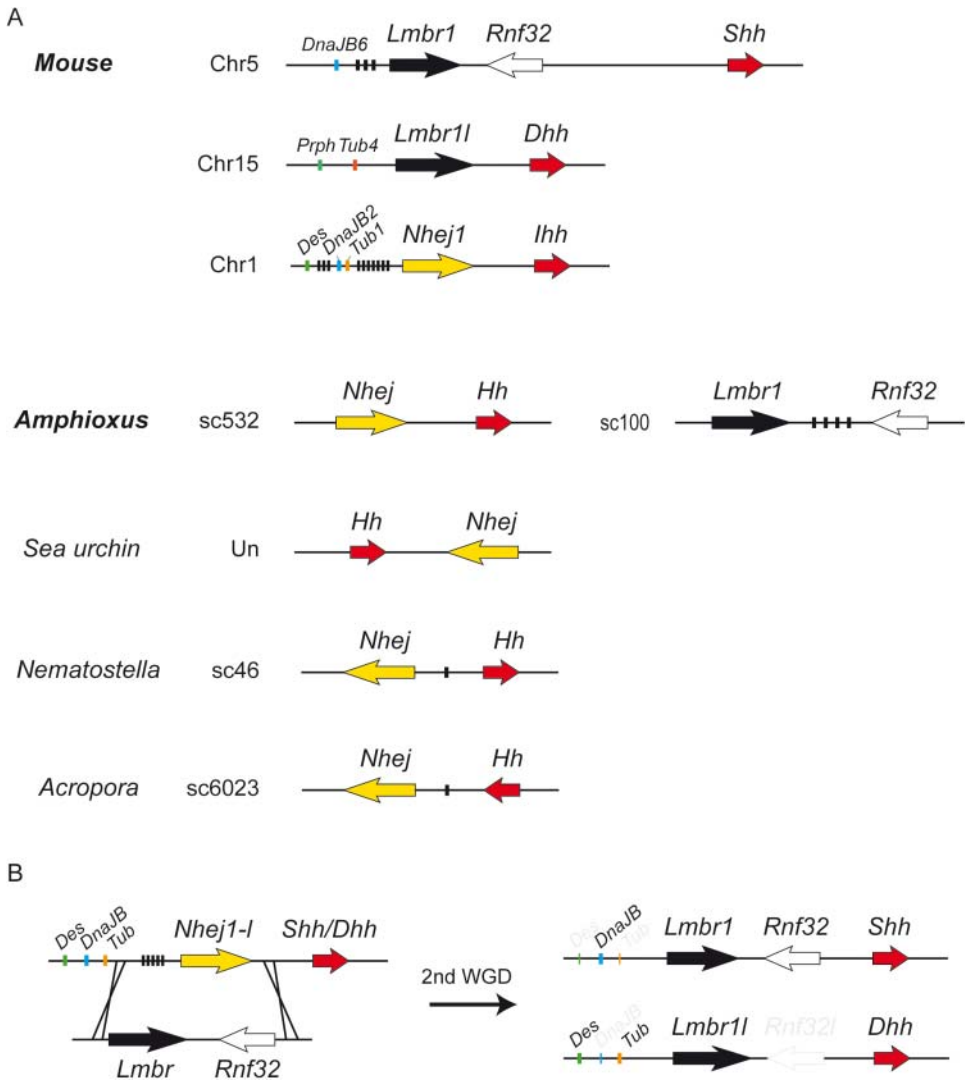

Figure 5 | Syntenic organization of *Hedgehog* genes and vertebrate-specific genomic rearrangement. (A) Genomic organization of the three *Hedgehog* paralogs in mouse (*Shh*, *Dhh* and *Ihh*) and in different selected invertebrates (red arrows, indicating the orientation of transcription). *Lmbr1*/*Nhej1*/*Rnf32* orthologs are represented by black/yellow/white arrows, respectively. Vertical bars represent intervening genes: green (*Prph*/*Des*), blue (*DnaJB2*/*DnaJB6*), orange (*Tub1*/*Tub4*) and black (other genes). Chromosome or scaffold is indicated for each species. (B) Possible evolutionary scenario for the insertion of the genomic fragment containing *Lmbr1* and *Rnf32* into the ancestral *Shh*/*Dhh* regulatory locus, some time between the two rounds of vertebrate WGD. *Nnej1* may have been lost along with the insertion or in a different event.

the region upstream of *Ihh* contains neither gene, but instead the phylogenetically unrelated *Nhej1* gene.

To determine the source of this discrepancy we then studied the chromosomal regions containing the single-copy ancestral *Hh* genes in invertebrates. We found that *Nhej1* is widely linked to *Hh* in invertebrates (amphioxus, sea urchin, sea anemone and the coral *Acropora*, Figure 5A), demonstrating that *Njeh1* was ancestrally linked to *Hh*, but has been lost in the *Shh* and *Dhh* regions. Interestingly, we found that neither *Lmbr1* nor *Rnf32* is linked to *Hh* in invertebrates; however, these genes are found linked to each other and with the same relative orientation in a different amphioxus genomic scaffold. Therefore, the simplest explanation for these data is that a small-scale rearrangement during vertebrate evolution introduced a chromosomal fragment including both *Lmbr1* and *Rnf32* into a (largely) intact *Hh* locus and likely removed another fragment containing a *Nhej1* paralog, creating the novel arrangement observed in *Shh* and *Dhh* (the latter of which has apparently subsequently lost *Rnf32*). Furthermore, that *Lmbr1* genes are found in the regions surrounding *Shh* and *Dhh*, but not in the third paralog *Ihh*, which in turn maintains the ancestral linkage to *Nhej1* present in invertebrate species, suggests that this arrangement arose after the first round of genome duplication (giving rise to the *Ihh* locus and the ancestor of the *Shh/Dhh* locus), but before the second duplication that gave rise to the separate *Shh* and *Dhh* loci, thus providing a very precise time point for origin of the now-key bystander relationship of *Shh* and *Lmbr1* and *Rnf32*: between the two ancestral vertebrate WGDs.

Discussion

Using comparative genomics and transgenesis in zebrafish we have investigated the evolution of the *Hedgehog* regulatory landscape within chordates. Despite remarkably conserved expression patterns during early embryonic development, we found little evidence for *cis*-regulatory conservation between the cephalochordate amphioxus and vertebrates, notwithstanding large conservation of non-coding regions within each lineage. In addition, we identified a vertebrate-specific genomic rearrangement, further differentiating the regulatory landscapes in both lineages.

Many *cis*-regulatory elements of *Shh* have been identified or defined by comparison of non-coding sequences among vertebrates[4,5,14,15,18–21,23], suggesting that *Shh* regulation is largely con-

served within the vertebrate lineage (or at least within tetrapods), with some elements, such as the limb enhancer MFCS1/ZRS, dating back to the origin of gnathostomes[4,22,44], and others of all vertebrates[40,45]. Similarly, we found widespread conservation of non-coding sequences among the three studied amphioxus species, spanning ~100 million years of cephalochordate evolution[36,38], at a level comparable to other loci with well-known conserved expression patterns (e.g. the *Hox* cluster[46]). On the other hand, we found no conservation of non-coding sequences between cephalochordates and vertebrates. Although the large evolutionary distance between both lineages (at least 535–550 my[47]) has often rendered too large for identification of conserved non-coding sequences, some HCNRs and cryptic conserved regulatory elements have been indeed identified for some important developmental genes with conserved expression patterns[41,42,48–50], suggesting that the regulation of *Hedgehog* loci is, at least, not particularly constrained over long evolutionary times compared to other genes with similarly crucial functions.

Further supporting the idea that the *Hedgehog* locus may have experienced large evolutionary divergence during chordate evolution, we also found no clear cases of cryptic conservation of regulatory elements in our transgenic assays. Only one out of the seven fragments (F) – spanning >30 Kbp of the amphioxus *Hh* locus – tested for enhancer activity in zebrafish drove reporter expression consistent with the endogenous zebrafish and amphioxus *Hedgehog* genes. This fragment overlaps ~0.5 Kbp with fragment G (Figure 4A), partly including a highly conserved block described above, but not the short conserved stretch reported by Rétaux *et al*[40]. Since fragment G did not drive similar expression, the reported enhancer activity may therefore lay within the upstream half of the second intron (red in Figure 4), and not within the largest highly conserved region. Importantly, this sequence promotes expression to pharyngeal endoderm and developing branchial arches. The only *Shh* enhancer with similar activity described to date[4], MRCS1 (Figure 1), does not lay within the orthologous intron, but far upstream, more than 500 Kbp away in mouse, and close to the bystander *Rnf32*, and it is conserved only from mammals to reptiles. In addition, in both mouse and zebrafish, extensive probing of both intron sequences for enhancer activity[15,18,22,51] did not show any equivalent enhancer in any of the two vertebrate species. This evolutionary divergence is also consistent with comparisons of *cis*-regulatory elements between mouse and zebrafish. Although some of the enhancers can be identified as orthologous by sequence similarity in the two vertebrate species, they hardly drive similar expression patterns when tested in reporter assays[22]. Therefore, *Shh* regulatory landscapes do not seem to be tightly constrained at the sequence level even within major vertebrate groups, despite the extensive expression pattern conservation observed across lineages.

However, it is important to note that several experimental limitations may lead to false negatives when probing sequences for regulatory activity. First, the amphioxus sequences are being tested in heterologous systems, not in their endogenous regulatory environments. Although amphioxus sequences have been extensively reported to be active in vertebrate systems[3,41,42,48–50], it is still unclear how sensitive and reliable the heterologous approach is. Second, in the specific case of the various *Shh* floor plate enhancers both in mouse and in zebrafish, they have been shown to be often codependent and their activity enhanced in a cooperative or synergic way[18,22,39]. Therefore, the combination of different amphioxus sequences could also be necessary to drive significant reporter expression. Unfortunately, this issue is very difficult to evaluate without knowing were the specific regulatory elements reside in amphioxus. Third, it is also possible that other amphioxus enhancers lay in further upstream or downstream regions, or even within the neighboring gene *Nhej1*, as the previously reported endochondral bone *Ihh* enhancer[26]. Consistent with this possibility, several discrete HCNRs are detected within the two long introns of the amphioxus *Nhej1* gene, comparing *B. floridae* and *B. belcheri* (Figure S4). Finally, only early developmental stages have been probed in this study, and thus it is possible that shared regulatory inputs do exist for later stages of development; however, large conservation of expression patterns between *Shh* and amphioxus *Hh* is observed only at these early stages[29,30].

Perhaps the most exciting finding of this study is the vertebrate-specific genomic novelty associated with the origin of the *Shh-Rnf32-Lmbr1* genomic regulatory block. First, these results suggest that *Dhh* and *Shh* may be more phylogenetically related to each other than to *Ihh*, in contrast to previous phylogenetic analyses[32,45,52], likely affected by the faster evolutionary rates of *Dhh* coding sequence. Second, this genomic novelty may be associated with a key novel expression domain of *Shh*. *Shh* is expressed in the limbs of all jawed vertebrates, including both bony and cartilaginous lineages[24,44]. The recruitment of *Hedgehog* signaling to these structures has been suggested as one of the crucial events for the origin of paired appendages, probably through the cooption of genetic programs that were already operating in the median fins[44,53]. Importantly, despite extensive searches for regulatory elements in different species, only one enhancer responsible for the limb expression of *Shh* has been identified to date, the MFCS1/ZRS enhancer, which is located within the fifth intron of the bystander gene *Lmbr1*[21] and is highly conserved across gnatostomate species[4,14,20,44]. Remarkably, our results demonstrate that the recruitment of the *Lmbr1* gene into the *Shh* regulatory landscape to establish a new GRB – and seemingly replace the old one integrated by *Nhej1*[26] – occurred within the vertebrate lineage, though a genome rearrangement between the two rounds of WGD. Whether *Lmbr1* already contained regulatory elements at the time of the genomic rearrangement or it simply provided the appropriate raw material for the evolution of the enhancer, this new syntenic configuration may have allowed the recruitment of *Shh* expression to the limbs. Although it may not be possible to confidently establish a causal relationship between the two evolutionary events, it suggests the exciting possibility that, in some cases, the remodeling of the genome architecture may underlay the evolution of gene regulation and the appearance of novel traits.

Methods

Genomic library screening and PCR-based cloning. We screened a Lambda Fix II/XhoI genomic library (Stratagene) of *B. lanceolatum*[54] with [α-32P] dCTP-labeled probes by random-hexamer priming. Approximately 6×10^5 recombinant phages were screened at standard conditions (60°C)[54]. For the primary screening, we used a probe for each of the three *B. lanceolatum* exons (EU754743). This strategy allowed the identification of positive phages containing the first (λ5020A), the second (λ5010B) and the second and third exons together (λ5009BC) and neighboring non-coding regions (Figure 2). We performed a second screening using a probe designed at the 5′ of λ5020A that provided 15 Kbp upstream the ATG (λ5032). All phages were sequenced by randomly interspersed primer-binding sites technology using a Tn7 transposon-based system (GPS®-1 Genome Priming System, New England BioLabs) and specific 'walking' primers, and the assembly was made by Phred, Phrap, and Consed software[55–57].

We next used a different strategy to clone further upstream *B. lanceolatum* genomic sequence. Taking advantage of the high polymorphism in the *B. floridae* amphioxus genome, we aligned the genomic sequences from the two *Hh* haplotypes (scaffolds 137 and 532) and selected blocks that had >99% conservation over long sequence stretches (300–600 bp). We then designed 2–3 forward and reverse primers spanning these regions and use them together in a single PCR reaction for each block using *B. lanceolatum* genomic DNA and low annealing temperature. We cloned and sequence the PCR products for each block using pCRII/TOPO vectors (Invitrogen). Then, between each block we designed *B. lanceolatum* specific primers and performed PCR reactions using iProof DNA polymerase (Promega) to amplify long fragments and cloned them. Using this strategy we cloned three new blocks, BS2-4 (Figure 2), that were sequenced using primers specifically designed for sequence walking. All primer sequences are available upon request. The whole assembled *Hh* locus from *B. lanceolatum* has been submitted to GeneBank (accession number JX034725).

Phylogenetic footprinting analyses. We used the following genomic sequences and annotations: (i) *B. floridae*, scaffold 532 combined with 137 when necessary, from *Nhej* (inclusive) to 2 Kbp downstream *Hh* (total ~82 Kbp); (ii) access to unpublished *B. belcheri* genome sequence was kindly provided by Dr. Anlong Xu, and the equivalent region to *B. floridae* was used (~68.5 Kbp); for vertebrates, the regions

including the bystander genes (*Lmbr1*/*Lmbr1l*, *Rnf32* and/or *Nhej1*) to the *Hedgehog* paralogs were extracted from Ensembl (see below), together with prebuilt VISTA annotations. Annotations for the amphioxus genes were done by sequence conservation to the *B. lanceolatum* orthologs (*BlHh*, EU754743; *BlNhej1*, JX034724). Orthology relationships could be unambiguously determined by best reciprocal blasts (see also below for *Lmbr1*).

Phylogenetic footprinting was performed using the visualization tool mVISTA[58] for multi-species alignments generated using the LAGAN software[59] (visualization of alignments produced by AVID and Shuffled-LAGAN yielded similar results). For comparisons between the three amphioxus species we used high stringency conditions for peak calling in the VISTA plots (window size of 300 bp, minimum width of 300 bp and sequence identity threshold of 80%). For comparisons within vertebrate paralogs and between vertebrates and amphioxus, we used standard conditions (window size of 100 bp, minimum width of 100 bp and sequence identity threshold of 70%) for vertebrates and lower stringent conditions for amphioxus (window size of 50 bp, minimum width of 50 bp and sequence identity threshold of 60%). Usage of different reference genomes for alignment visualization (i.e. amphioxus, medaka or Xenopus) yielded similar results.

Synteny comparisons and genomic resources. We used the following genome resources to browse and search for orthologs of *Hh, Nhej1, Lmbr1* and *Rnf32*: *Trichoplax adhaerens* Grell-BS-1999 v1.0, *Nematostella vectensis* v1.0, *B. floridae* v1.0, *Ciona intestinalis* v2.0 and v1.0, *Daphnia pulex* v1.0, *Lottia gigantea* v1.0 and *Capitella teleta* v1.0, at DOE Joint Genome Institute (JGI) webpage (http://genome.jgi-psf.org/euk_home.html), and of *Strongylocentrotus purpuratus* Build 2.1, *Drosophila melanogaster* Build Fb5.3, *Homo sapiens* Build GRCh37, *Mus musculus* Build 37.1, *Gallus gallus v2.1, Anolis carolinensis* AnoCar1.0, *Xenopus tropicalis* JGI v4.1, and *Oryzias latipes* at the Ensembl webpage (http://www.ensembl.org), and *Saccoglossus kowalevskii* 2008-Dec-09 scaffolds at HGSC Baylor College of Medicine webpage (http://blast.hgsc.bcm.tmc.edu), and the *Acropora digitifera* genome[60]. Paralogs for the different *Shh*, *Dhh* and *Ihh* neighbouring genes in vertebrates were obtained by the Ensembl paralog tool. Lamprey could not be included in the analyses because of the current incomplete and fragmentary genomic assembly, in particular for both *Hh* genes[45].

Phylogenetic analyses of Lmbr1/Lmbr1l genes. We downloaded full protein sequences for *Lmbr1* and *Lmbr1l* from *H. sapiens*, *X. tropicalis* and *D. rerio,* and *Lmbr1* orthologs from the following invertebrates: *B. floridae*, *B. belcheri, S. purpuratus*, *S. kowalevskii*, *L. gigantea*, *C teleta*, *D. melanogaster, D. pulex*, *N. vectensis* from the sources mentioned above. Sequences for *Apis mellifera* and *Tribolium castaneum* were obtained through NCBI BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi). In addition we used putative *Lmbrd1* genes from *H. sapiens*, *B. floridae*, *L. gigantea* and *N. vectensis* as outgroups. Phylogenetic trees were generated by the Bayesian method with MrBayes 3.1.2[61,62] using two independent runs (each with four chains). Model selection using ProtTest[63] (best model: CpRev+G), convergence determination, burn-in, and consensus tree calculations were done as previously described[64,65]. In total, 3,000,000 generations were run, reaching convergence at generation 685,000; all trees prior convergence were discarded, and the remaining ones were used to build the consensus tree (Figure S5). This tree shows that all investigated *Lmbr1* genes are orthologs and that *Lmbr1* and *Lmbr1l* in vertebrates arose from a vertebrate-specific duplication, most likely one of the two WGDs.

Transgenic analyses in zebrafish and in situ hybridization. We designed primers to amplify seven overlapping *B. lanceolatum* genomic fragments of 4–7.0 Kbp. PCRs were performed on *B. lanceolatum* genomic DNA or on the corresponding phage DNA extractions using iProof™ High-Fidelity DNA Polymerase (Bio-Rad). PCR products were cloned in pCR8GW/TOPO vector (Invitrogen) according to manufacturer. Sequence-verified clones were then transferred with the Gateway recombination system (Invitrogen) to the ZED vector[66]. The final transgenic constructs were purified using phenol-chlorophorm and normalized at 50 ng/ml in DEPC water prior to microinjection. For each construct, >100 injected embryos were assayed and GFP expression investigated at 24 and 48 hpf. RFP expression within the muscles observed 72 hpf served as a positive control for transgenesis. Two constructs showed consistent GFP expression in F0 (D, 28/130 injected embryos (22%) and F, 25/165 (15%)); of these, only F – which showed an expression pattern consistent with the endogenous *shha* gene – was raised to the next generation to obtain stable transgenic lines (F1 lines), in addition to fragments D and G, with negative GFP activity. For embryonic gene expression analysis of GFP driven by fragment F and endogenous *shha* by *in situ* hybridization, zebrafish embryos were fixed at different stages in 4% paraformaldehyde overnight at 4°C; and *in situ* hybridizations carried out as previously described[67].

1. Engstrom, P. G., Ho Sui, S. J., Drivenes, O., Becker, T. S. & Lenhard, B. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome Res.* **17**, 1898–1908, doi:10.1101/gr.6669607 (2007).
2. Kikuta, H. *et al.* Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res.* **17**, 545–555 (2007).
3. Maeso, I. *et al.* An ancient genomic regulatory block conserved across bilaterians and its dismantling in tetrapods by retrogene replacement. *Genome Res* **22**, 642–655 (2012).
4. Sagai, T. *et al.* A cluster of three long-range enhancers directs regional Shh expression in the epithelial linings. *Development* **136**, 1665–1674 (2009).
5. Lettice, L. A. *et al.* Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. *Proc Natl Acad Sci USA* **99**, 7548–7553 (2002).
6. Varjosalo, M. & Taipale, J. Hedgehog: functions and mechanisms. *Genes Dev* **22**, 2454–2472 (2008).
7. Ingham, P. W. & McMahon, A. P. Hedgehog signaling in animal development: paradigms and principles. *Genes Dev* **15**, 3059–3087 (2001).
8. Wilson, L. & Maden, M. The mechanisms of dorsoventral patterning in the vertebrate neural tube. *Dev Biol* **282**, 1–13 (2005).
9. Kiecker, C. & Lumsden, A. Hedgehog signaling from the ZLI regulates diencephalic regional identity. *Nature Neuroscience* **7**, 1242–1249 (2004).
10. Riddle, R. D., Johnson, R. L., Laufer, E. & Tabin, C. Sonic hedgehog mediates the polarizing activity of the ZPA. *Cell* **75**, 1401–1416 (1993).
11. Brown, A. S. & Epstein, D. J. Otic ablation of smoothened reveals direct and indirect requirements for Hedgehog signaling in inner ear development. *Dev Cell* **22**, 585–596 (2012).
12. Chuong, C. M., Patel, N., Lin, J., Jung, H. S. & Widelitz, R. B. Sonic hedgehog signaling pathway in vertebrate epithelial appendage morphogenesis: perspectives in development and evolution. *Cell Mol Life Sci* **57**, 1672–1681 (2000).
13. Echelard, Y. *et al.* Sonic hedgehog, a member of a family of putative signaling molecules, is implicated in the regulation of CNS polarity. *Cell* **75**, 1417–1430 (1993).
14. Goode, D. K., Snell, P., Smith, S. F., Cooke, J. E. & Elgar, G. Highly conserved regulatory elements around the SHH gene may contribute to the maintenance of conserved synteny across human chromosome 7q36.3. *Genomics* **86**, 172–181 (2005).
15. Epstein, D. J., McMahon, A. P. & Joyner, A. L. Regionalization of Sonic hedgehog transcription along the anteroposterior axis of the mouse central nervous system is regulated by Hnf3-dependent and -independent mechanisms. *Development* **126**, 281–292 (1999).
16. Martí, E., Takada, R., Bumcrot, D. A., Sasaki, H. & McMahon, A. P. Distribution of Sonic hedgehog peptides in the developing chick and mouse embryo. *Development* **121**, 2537–2547 (1995).
17. Bardet, S. M., Ferran, J. L., Sanchez-Arrones, L. & Puelles, L. Ontogenetic expression of sonic hedgehog in the chicken subpallium. *Front Neuroanat* **4. pii**, 28 (2010).
18. Jeong, Y., El-Jaick, K., Roessler, E., Muenke, M. & Epstein, D. J. A functional screen for sonic hedgehog regulatory elements across a 1 Mb interval identifies long-range ventral forebrain enhancers. *Development* **133**, 761–772 (2006).
19. Sagai, T., Hosoya, M., Mizushina, Y., Tamura, M. & Shiroishi, T. Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. *Development* **132**, 797–803 (2005).
20. Sagai, T. *et al.* Phylogenetic conservation of a limb-specific, cis-acting regulator of Sonic hedgehog (Shh). *Mammalian Genome* **V15**, 23–34 (2004).
21. Lettice, L. A. *et al.* A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* **12**, 1725–1735 (2003).
22. Ertzer, R. *et al.* Cooperation of sonic hedgehog enhancers in midline expression. *Developmental Biology* **301**, 578–589 (2007).
23. Lang, M. *et al.* Conservation of shh cis-regulatory architecture of the coelacanth is consistent with its ancestral phylogenetic position. *EvoDevo* **1**, 11 (2010).
24. Davis, M. C., Dahn, R. D. & Shubin, N. H. An autopodial-like pattern of Hox expression in the fins of a basal actinopterygian fish. *Nature* **447**, 473–476 (2007).
25. Putnam, N. *et al.* The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**, 1064–1071 (2008).
26. Klopocki, E. *et al.* Copy-number variations involving the IHH locus are associated with syndactyly and craniosynostosis. *Am J Hum Genet* **88**, 70–75 (2011).
27. Pani, A. M. *et al.* Ancient deuterostome origins of vertebrate brain signalling centres. *Nature* **483**, 289–294 (2012).
28. Nüsslein-Volhard, C. & Wieschaus, E. Mutations affecting segment number and polarity in Drosophila. *Nature* **287**, 795–801 (1980).
29. Rink, J. C., Gurley, K. A., Elliott, S. A. & Sánchez Alvarado, A. Planarian Hh signaling regulates regeneration polarity and links Hh pathway evolution to cilia. *Science* **326**, 1406–1410 (2009).
30. Seaver, E. C. & Kaneshige, L. M. Expression of 'segmentation' genes during larval and juvenile development in the polychaetes Capitella sp. I and H. elegans. *Dev Biol* **289**, 179–194 (2006).
31. Walton, K. D., Warner, J., Hertzler, P. H. & McClay, D. R. Hedgehog signaling patterns mesoderm in the sea urchin. *Dev Biol* **331**, 26–37 (2009).
32. Shimeld, S. M. The evolution of the hedgehog gene family in chordates: insights from amphioxus hedgehog. *Development Genes and Evolution* **209**, 40–47 (1999).
33. Shimeld, S. M., van den Heuvel, M., Dawber, R. & Briscoe, J. An Amphioxus Gli Gene Reveals Conservation of Midline Patterning and the Evolution of Hedgehog Signalling Diversity in Chordates. *PLoS ONE* **2**, e864 (2007).
34. Irimia, M. *et al.* Conserved developmental expression of Fezf in chordates and Drosophila and the origin of the Zona Limitans Intrathalamica (ZLI) brain organizer. *EvoDevo* **1**, 7 (2010).
35. Osorio, J., Mazan, S. & Retaux, S. Organisation of the lamprey (Lampetra fluviatilis) embryonic brain: Insights from LIM-homeodomain, Pax and hedgehog genes. *Dev Biol* **288**, 100–112 (2005).

36. Kon, T. *et al.* Phylogenetic position of a whale-fall lancelet (Cephalochordata) inferred from whole mitochondrial genome sequences. *BMC Evol Biol* **7**, 127 (2007).
37. Nohara, M., Nishida, M., Miya, M. & Nishikawa, T. Evolution of the mitochondrial genome in cephalochordata as inferred from complete nucleotide sequences from two epigonichthys species. *J Mol Evol* **60**, 526–537 (2005).
38. Nohara, M., Nishida, M. & Nishikawa, T. New complete mitochondrial DNA sequence of the lancelet Branchiostoma lanceolatum (Cephalochordata) and the identity of this species' sequences. *Zoolog. Sci.* **22**, 671–674 (2005).
39. Jeong, Y. & Epstein, D. J. Distinct regulators of Shh transcription in the floor plate and notochord indicate separate origins for these tissues in the mouse node. *Development* **130**, 3891–3902 (2003).
40. Rétaux, S. & Kano, S. Midline signaling and evolution of the forebrain in chordates: a focus on the lamprey Hedgehog case. *Integr Comp Biol* **50**, 98–109 (2010).
41. Holland, L. Z. *et al.* The amphioxus genome illuminates vertebrate origins and cephalochordate biology. *Genome Res* **18**, 1100–1111 (2008).
42. Royo, J. L. *et al.* Transphyletic conservation of developmental regulatory state in animal evolution. *Proc Natl Acad Sci USA* **108**, 14186–14191 (2011).
43. Strähle, U., Blader, P. & Ingham, P. W. Expression of axial and sonic hedgehog in wildtype and midline defective zebrafish embryos. *Int J Dev Biol* **40**, 429–440 (1996).
44. Dahn, R. D., Davis, M. C., Pappano, W. N. & Shubin, N. H. Sonic hedgehog function in chondrichthyan fins and the evolution of appendage patterning. *Nature* **445**, 311–314 (2007).
45. Kano, S. *et al.* Two lamprey Hedgehog genes share non-coding regulatory sequences and expression patterns with gnathostome Hedgehogs. *PLoS One* **5**, e13332 (2010).
46. Pascual-Anaya, J., D'Aniello, S. & Garcia-Fernàndez, J. Unexpectedly large number of conserved noncoding regions within the ancestral chordate Hox cluster. *Dev Genes Evol* **218**, 591–597 (2008).
47. Satoh, N. The ascidian tadpole larva: comparative molecular development and genomics. *Nat Rev Genet* **4**, 285–295 (2003).
48. Punnamoottil, B. *et al.* Cis-regulatory characterization of sequence conservation surrounding the Hox4 genes. *Dev Biol* **340**, 269–282 (2010).
49. Manzanares, M. *et al.* Conservation and elaboration of Hox gene regulation during evolution of the vertebrate head. *Nature* **408**, 854–857 (2000).
50. Hufton, A. L. *et al.* Deeply conserved chordate noncoding sequences preserve genome synteny but do not drive gene duplicate retention. *Genome Res* **19**, 2036–2051 (2009).
51. Muller, F. *et al.* Intronic enhancers control expression of zebrafish sonic hedgehog in floor plate and notochord. *Development* **126**, 2103–2116 (1999).
52. Matus, D. Q., Magie, C. R., Pang, K., Martindale, M. Q. & Thomsen, G. H. The Hedgehog gene family of the cnidarian, Nematostella vectensis, and implications for understanding metazoan Hedgehog pathway evolution. *Dev Biol* **313**, 501–518 (2008).
53. Freitas, R., Zhang, G. & Cohn, M. J. Evidence that mechanisms of fin development evolved in the midline of early vertebrates. *Nature* **442**, 1033–1037 (2006).
54. Cañestro, C. *et al.* Amphioxus alcohol dehydrogenase is a class 3 form of single type and of structural conservation but with unique developmental expression. *Eur J Biochem* **267**, 6511–6518 (2000).
55. Ewing, B., Hillier, L., Wendl, M. C. & Green, P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8**, 175–185 (1998).
56. Gordon, D., Abajian, C. & Green, P. Consed: a graphical tool for sequence finishing. *Genome Res* **8**, 195–202 (1998).
57. Ewing, B. & Green, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8**, 186–194 (1998).
58. Mayor, C. *et al.* VISTA : visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16**, 1046–1047 (2000).
59. Brudno, M. *et al.* LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res* **13**, 721–731 (2003).
60. Shinzato, C. *et al.* Using the Acropora digitifera genome to understand coral responses to environmental change. *Nature* **476**, 320–323 (2011).
61. Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755 (2001).
62. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574 (2003).
63. Abascal, F., Zardoya, R. & Posada, D. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* **21**, 2104–2105 (2005).
64. Irimia, M. *et al.* Contrasting 5′ and 3′ Evolutionary Histories and Frequent Evolutionary Convergence in Meis/hth Gene Structures. *Genome Biol Evol* **3**, 551–564 (2011).
65. D'Aniello, S. *et al.* Gene expansion and retention leads to a diverse tyrosine kinase superfamily in amphioxus. *Mol Biol Evol* **25**, 1841–1854 (2008).
66. Bessa, J. *et al.* Zebrafish enhancer detection (ZED) vector: a new tool to facilitate transgenesis and the functional analysis of cis-regulatory regions in zebrafish. *Dev Dyn* **238**, 2409–2417 (2009).
67. Tena, J. J. *et al.* Odd-skipped genes encode repressors that control kidney development. *Dev Biol* **301**, 518–531 (2007).

Acknowledgements

We thank Dr. Anlong Xu for kindly proving access to the unpublished genome sequence of *B. belcheri*, and Scott W. Roy for invaluable help on the microsynteny section, and Renata Freitas and Isabel Almudí for helpful comments on the manuscript. M.I. wishes to thank Douglas Epstein for helpful conversations and advise. M.I., D.B., I.M. and J.G.-F. were funded by Grants BFU2005-00252 and BMC2008-03776 and BMC2011-23291 from the Spanish Ministry of Science and Innovation, and J.G.F. by the ICREA Academia Prize. M.I., D.B. and I.M. held FPI, APIF-UB and FPU fellowships, respectively. J.-L.G.-S. and J.L.R. were supported by Grants BFU2010-14839, CSD2007-00008875 (MEC), and CVI 3488 (Junta de Andalucía). JLR holds a JAE-Doc grant from the National Research Council, founded by Social European Funds.

Author contributions

MI conceived and designed the study; MI performed the cloning and sequencing of the *B. lanceolatum Hh* locus; MI and IM carried out the bioinformatic analyses; JLRP and JLGS performed the zebrafish transgenesis; DB and MI cloned the amphioxus sequences for transgenesis; MI, IM, JLGS and JGF wrote the manuscript and coordinated the analyses.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Irimia, M. *et al.* Comparative genomics of the *Hedgehog* loci in chordates and the origins of *Shh* regulatory novelties. *Sci. Rep.* **2**, 433; DOI:10.1038/srep00433 (2012).